To appear in the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK

LEARNING TO FUSE LATENT REPRESENTATIONS FOR MULTIMODAL DATA

Oyebade K. Oyedotun, Djamila Aouada, Björn Ottersten

Interdisciplinary Centre for Security, Reliability and Trust (SnT),
University of Luxembourg, L-1855, Luxembourg
{oyebade.oyedotun, djamila.aouada, bjorn.ottersten}@uni.lu

ABSTRACT

Multimodal learning leverages data from different modalities to improve the performance of a trained model. Typically, latent representations extracted from multimodal data are provided via direct feature fusion for end-to-end training of a deep neural network towards a specific task. However, the informativeness of the different data modalities can easily vary across a collected dataset. As such, *naively* or directly fusing the latent representations obtained for one modality and the other, as is commonly done in state-of-the-art works, may burden the model in finding concise representations that are indeed useful for learning. In this paper, we propose to instead learn the fusion of latent representations for multimodal data by using a modality gating mechanism that allows the dynamic weighting of extracted latent representations based on their informativness. Extensive experiments using the BU-3DFE dataset for facial expression recognition and the Washington object classification multimodal RGB-D dataset show that learning the fusion of the latent representations for different data modalities leads to improved model generalization than the conventional naive fusion method.

Index Terms— Multimodal learning, deep neural network, latent representation, fusion, classification

1. INTRODUCTION

Several computer vision tasks benefit from the powerful representation capacity of deep neural networks (DNNs). The latent representations obtained from trained DNNs typically capture the different factors of variations in the training data [1, 2]. DNN features obtained in the earlier layers of the model are usually low-level features characterizing input data attributes such as edges at different orientations and corners. The features extracted in the later layers are generally more abstract (or high-level) for performing different tasks such as object classification, segmentation, scene understanding, etc. Generally, DNNs are trained using inputs of a single modality. For example, RGB images [3], grayscale images [4], binary images [2], audio signal waveforms [5], etc. However, different works [6, 7, 8, 9] have proposed learning DNNs with multimodal data; such works show that multimodal data can be leveraged to improve the robustness of the latent representations learned towards better model generalization. In particular, the aforementioned works show that DNNs trained using multimodal data often generalize better as compared to similar DNNs trained on only one data modality. In [6], a dataset of RGB and images were collected for object classification. A DNN architecture based on pre-trained AlexNet was constructed for concurrent processing of RGB and depth images. Subsequently, the extracted latent representations obtained from the pre-trained AlexNet were directly fused via direct concatenation for end-to-end training of the whole model. Reported results show that learning from both RGB and depth data modalities yielded better results as opposed to learning from either of the two modalities singly. In another work [7], image and audio data modalities were leveraged for improving the performance of DNN on emotion recognition. Convolutional neural network (CNN) was used for extracting latent representations for image data, while a Deep Belief Network (DBN) was used for learning latent representa-

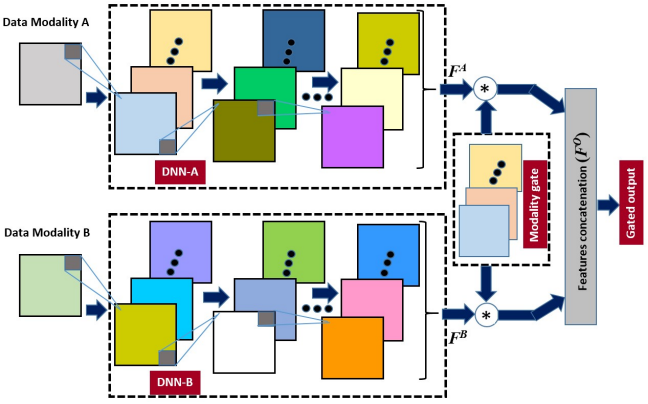

Fig. 1. Proposed data dependent fusion of latent representations, F^A and F^B extracted using DNN-A and DNN-B respectively. A modality gate is used for filtering F^A and F^B to realize improved fusion

This work was funded by the National Research Fund (FNR), Luxembourg, under the project references R-AGR-0424-05-D/Björn Ottersten and CPPP17/ IS/11643091/IDform/Aouada.

tions for processed audio data based on Mel-frequency cepstral coefficients (MFCCs). Although multimodal learning with naive fusion[1] (NF) works, one problem is that the level of informativeness of the different data modalities are typically ignored for learning. Existing works basically concatenate extracted latent representations of different data modalities and feed them into a stack of DNN layers irrespetive of the fact that one data modality may be far less informative (or corrupted) than the other; this can for instance distort learning. The assumption is that layers following the fusion stage in the DNN can somehow manage to extract decent latent representations for the classification layer. However, we argue that instead of burdening the DNN to cope with the aforementioned problem, employing a DNN that is specifically designed to address this scenario would lead to a more concise learning and model generalization. In extreme cases, given a pair of multimodal inputs for training, one of the modalities may be so noisy as to even *corrupt* learning given the input from the other data modality that truely characterizes the underlying factor of variation in the problem being modeled.

In this paper, we propose to learn how to fuse the different features extracted from multimodal data as opposed to direct fusion. We extend our recent work [10] to multimodal learning. In [10], we proposed a highway block for gating transformed and untransformed features in very deep networks, and as such alleviates the problem of model optimization. First, we clarify that the purpose of using a similar gating mechanism in this paper is not for addressing the problem of training very deep networks as in [10, 11]. In contrast, such a gating mechanism with appropriate modifications can be used to learn the fusion of latent representations for improved model generalization as presented in this paper. We use RGB and depth modalities to experimentally validate the proposed approach. Specifically, the proposed fusion approach offers the following contributions:

1. Learn appropriate fusion (weighting) of RGB and depth latent representations; this alleviates learning problems that can result from inconsistent informativeness of RBG and depth data over collected datasets.

2. Realize improved model generalization as opposed to naive fusion of latent representations using BU-3D facial expression recognition and Washington object classification RGB-D datasets.

The remainder of this paper is organized as follows. In Section 2, background and problem statement are presented. Section 3 gives the details of the proposed approach for learning the fusion of RBG and depth latent representations. Section 4 reports our experimental results. We conclude the paper in Section 5.

[1]Naive fusion and direct fusion are used interchangeably

2. BACKGROUND AND PROBLEM STATEMENT

2.1. Background on learning with multimodal data

Unimodal data can often have many training samples which lack fine details or that are even unreliable for capturing reasonable latent representations, due to noise, occlusion, illumination changes, etc. This can negatively impact learning of consistent latent representations, and consequently limit the generalization capacity of a trained model. Hence, learning from multimodal data can improve the consistency and reliability of learned latent representations; that is, where one data modality fails to be sufficiently informative, the other data modality can be leveraged to still render decent latent representations. In view of this, many works extract latent representations (e.g using pre-trained models) from multimodal data and thereafter perform direct fusion via concatenation to carry out end-to-end training [6, 7, 8, 9]. Assume a multimodal training dataset of the form

$$D = \{((x_n^{M_1}, x_n^{M_2}), y_n)\}_{n=1}^{N}, \tag{1}$$

where $x_n^{M_1}$ and $x_n^{M_2}$ are the n^{th} input from the data modality M_1 and the data modality M_2, respectively; y_n is the desired output for the n^{th} input data. The main assumption in multomodal learning is that $x_n^{M_1}$ and $x_n^{M_2}$ mostly provide complementary information for capturing the important factors of variations in a specific task such as classification, detection, regression, etc. Otherwise, multimodal learning may not yield any benefit in the case where both data modalities provide information for explaining the same factors of variations. For the rest of this paper, factors of variations are denoted using a set F. We use factors of variations to refer to the actual features that explain a specific task. For example, in a facial recognition task, features such as raised eyebrows, unelevated eyebrows, dropped jaw, closed mouth, slightly opened mouth, widely open mouth, etc. are different factors of variations that can explain facial expressions.

Let us denote the total factors of variations for a specific task S_T by F^T, and consider that they are *altogether* captured in a hypothetical dataset, T. Also, assume that we have a dataset A that partially captures F^T as F^A such that we can write[2] $F^A \subset F^T$; this is the case with unimodal data for learning. However, we can have multimodal data A and data B subject to $A \subset T$ and $B \subset T$ such that we can capture factors of variations F^A and F^B, respectively. Furthermore, if we encapsulate F^A and F^B using a multimodal dataset D as in (1), then we have $F^D = \{F^A, F^B\}$ and thus can write a general relation

$$F^A, F^B \subseteq F^D \subset F^T. \tag{2}$$

Note from (2) that mulitmodal learning becomes unuseful when $F^A \cup F^B = F^A$, since we have the relation

$$F^A = F^D \subset F^T. \tag{3}$$

[2]Since A is a snaphot of the task domain defined by T

Particularly, observe that this can happen even when data A and B are different in the input data space, since F^A and F^B actually depend on extracted latent representations. Therefore latent representation and factors of variations are used interchangeably.
Multimodal learning becomes useful when $F^A \cup F^B \neq F^A$; this is the goal of multimodal learning. Subsequently, we can write (2) as

$$F^A \subset F^D \subset F^T. \tag{4}$$

We posit that the level of contribution of modality data B for learning a specific task based on modality data A can be related to $F^B \backslash F^A = \{f \epsilon F^A \mid f \notin F^B\}$; specifically, considering its cardinality as in $\mid F^B \backslash F^A \mid$. Furthermore, we can write a general expression for F^T as

$$F^T = \{F^A \backslash F^B, F^B \backslash F^A, F^R\}, \tag{5}$$

where F^R contains the remaining factors of variations that can be captured using other datasets.

2.2. Problem statement

In many real-life scenarios, we will have that $F^A \backslash F^T \neq \emptyset$ and $F^B \backslash F^T \neq \emptyset$, since data $A \subseteq T$ and $B \subseteq T$ subject to $A \cup T \neq A$ and $B \cup T \neq B$, respectively. Therefore, we have that the factors of variations that are actually useful for task T denoted F_u^A and F_u^B are such that $\exists\ F_u^A \subset F^A$ and $F_u^B \subset F^B$; u is used to index a set of useful $f \in F$. As such, we can have the following cases:

- Where $\mid F_u^A \mid \approx \mid F_u^B \mid$, and naive fusion looks interesting, since subsequent layers in the model can *almost equally* rely on F_u^A and F_u^B for model generalization.
- Where $\mid F_u^A \mid \gg \mid F_u^B \mid$ or $\mid F_u^B \mid \gg \mid F_u^A \mid$, and performing a naive fusion may *burden* subsequently layers with focusing on what is more important, since the succeeding layers in the model are trying to learn F_u^A and F_u^B with *roughly equal importance* for generalization.

The second scenario is not uncommon in practice, and as such is addressed in this paper. The following section discusses the proposed approach taken to tackle the highlighted problem.

3. LEARNING TO FUSE RGB-DEPTH LATENT REPRESENTATIONS

For alleviating the problem mentioned in Section 3.2, we propose to realize a fusion scheme that is *data driven (i.e. dynamic)* and allows the model to learn by itself the importance (or weighting) of the different extracted latent representations, F^A and F^B. The goal is that the proposed *modality gate* filters F^A and F^B, and thereby passes mostly F_u^A and F_u^B to the features concatenation stage that completes the fusion phase as shown in Fig. 1; fully connected layers and a softmax layer can be trained on top of the fusion outcome. Namely, our approach in this paper for constructing the modality gate is motivated by [10, 11]; that is, using a mechanism for learning what part of the latent representations are routed to subsequent layers in the model. Again, it is emphasized that the purpose of using a gating mechanism in this paper is not to alleviate training problems of very deep network as in [10, 11]. Specifically, this paper uses a similar form of gating mechanism proposed in [10] which has improved learning characteristics over that in [11] for learning fusion of latent representations for multimodal data.
The modality gate used in this paper and shown in Fig. 1 is of the form

$$G = \varphi(WF^B + \theta), \tag{6}$$

where G denotes gate transformation on F using weight W, bias θ and Log-sigmoid activation function φ. G learns the part of F^B that is captured for fusion. Subsequently, $1 - G$ is used for selecting the part of F^B that is captured for fusion. The overall formulation of the modality gate given in Fig.1 can thus lead to the fusion outcome F^O as follows

$$F^O = \{F^A(1 - (G(F^B)) \otimes F^B(G(F^B)\}, \tag{7}$$

where $\otimes$ is the concatenation operation. Other notations remain as earlier stated. Following the fusion stage, additional weight layers can be stacked in the model and trained end-to-end using F^O as input.

4. EXPERIMENTS

In this section, experimental results to validate the usefulness of learning to fuse latent representations as opposed to naive fusion (NF) are reported. Namely, the BU-3DFE facial expression [12] and the Washington object RGB-D classification dataset [13] are used. For the constructed *RGB-ResNet50+DepthMap-scratch* model, F^A is extracted from RGB images using pre-trained ResNet50 and F^B is extracted from depth maps using a model trained from scratch[3]. For the proposed *RGB-VGG19+DepthMap-scratch* model, F^A is extracted from RGB images using pre-trained VGG19 and F^B is extracted from depth images using a model trained from scratch. Fusion via modality gating is performed using G as in (7), after which two fully connected and softmax layers are stacked on top of it as discussed in Section 3. The whole model is then trained end-to-end using Adaptive Moment (ADAM) mini-batch gradient descent. An initial learning rate of 10^{-3} is used; the learning is annealed to a final value of 10^{-5}. All models are trained for a maximum of 400 epochs.

4.1. BU-3DFE facial expression RGB-D dataset

BU-3DFE dataset contains both 3D and 2D data modalities for facial expressions of 100 different subjects. The data processing in [14] is followed for preparing the training data.

[3]Pre-trained model give poor results with depth maps as reported in [14]

Approach	Test acc. (%)
DepthMap-ResNet50 [14]	61.11
DepthMap-VGG19 [14]	28.06
DepthMap-scratch [14]	84.72
RGB-ResNet50 [14]	82.92
RGB-VGG19 [14]	81.25
NF: RGB-ResNet50+DepthMap-scratch	87.08
NF: RGB-VGG19+DepthMap-scratch	89.31
Ours: RGB-ResNet50+DepthMap-scratch	**89.86**
Ours: RGB-VGG19+DepthMap-scratch	**90.69**

Table 1. BU-3DFE RGB-D dataset results with 10-fold CV

Approach	Test acc. (%)
3D geometric shape model+LDA [15]	83.60
Bayesian Belief net+statistical facial features [16]	82.30
Distance+slopes+SVM [18]	87.10
2D+3D features fusion+SVM [19]	86.32
Geometric scattering representation+SVM [20]	84.80
Geometric+photometric attributes+VGG19 [21]	84.87
NF:RGB-ResNet50+DepthMap-scratch [14]	87.08
NF: RGB-VGG19+DepthMap-scratch [14]	89.31
Ours: RGB-ResNet50+DepthMap-scratch	**89.86**
Ours: RGB-VGG19+DepthMap-scratch	**90.69**

Table 2. Results comparison on BU-3DFE dataset

Also, a similar experimental setting is used for extracting latent expresentations from both data modalities; that is, using pre-trained models (ResNet-50 and VGG19) on ImageNet dataset for RGB data and training a DNN from scratch on depth data. Furthermore, 10-fold cross-validation (CV) is employed for evaluating the models as in [14, 15, 16, 17]. The results are given in Table 1, along with models trained on RGB only, depth map only and via naive fusion as in [14]. It will be seen that the proposed fusion approach gives improved results over naive fusion. Results comparison with state-of-the-art models that further validates our proposal is presented in Table 2.

4.2. Washington object RGB-D dataset

For Washington object classification dataset, we collect a total of 4500 samples for both 3D and 2D data modalities to *clearly* demonstrate that naive fusion can sometimes even hurt model performance as discussed in Section 3.2; the samples span the different 51 categories in the dataset. We perform 10-fold and 5-fold cross-validation (CV) using RGB only, depth only, naive fusion (NF) and the proposed fusion approach. Table 3 and Table 4 show obtained experimental results for 10-fold and 5-fold CV, respectively. Again, it will be observed that using depth maps with pre-trained models gives poor results; see Tables 3 & 4. Importantly, it is observed that naive fusion using RGB and depth modalities results in degraded model performance as compared to when either of the two data modalities are used. From Tables 3 & 4, *DepthMap-scratch* outperforms *NF:RGB-ResNet50+DepthMap-scratch*; and *RGB-VGG19* outperforms *NF:RGB-VGG19+DepthMap-scratch*. Interestingly, it is observed that learning the fusion stage gives improved results as compared to naive fusion. This shows that the proposed fusion approach can learn appropriate weightings for the different latent representations such that not only is performance degradation mitigated, but notably improved.

Approach	Test acc. (%)
Depth map-ResNet50	2.39
Depth map-VGG19	2.38
DepthMap-scratch	77.07
RGB-ResNet50	50.08
RGB-VGG19	98.25
NF: RGB-ResNet50+DepthMap-scratch	76.65
NF: RGB-VGG19+DepthMap-scratch	90.56
Ours: RGB-ResNet50+DepthMap-scratch	**99.32**
Ours: RGB-VGG19+DepthMap-scratch	**99.54**

Table 3. Washington RGB-D dataset results with 10-fold CV

Approach	Test acc. (%)
Depth map-ResNet50	1.67
Depth map-VGG19	2.38
DepthMap-scratch	73.87
RGB-ResNet50	29.76
RGB-VGG19	96.73
NF: RGB-ResNet50+DepthMap-scratch	61.23
NF: RGB-VGG19+DepthMap-scratch	83.16
Ours: RGB-ResNet50+DepthMap-scratch	**98.93**
Ours: RGB-VGG19+DepthMap-scratch	**99.51**

Table 4. Washington RGB-D dataset results with 5-fold CV

5. CONCLUSION

Multimodal learning is useful for improving model performance. The latent representations extracted from the different modalities are typically fused via direct concatenation. We argue and show that such fusion is naive and can even hurt model performance. In contrast, we propose in this paper to allow the model to learn a data driven fusion stage using a gate mechanism that filters latent representations from multimodal data. Experimental results on two different datasets validate the proposed fusion approach. Improved results over naive fusion and several state-of-the-art models are obtained.

6. REFERENCES

[1] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 808–822.

[2] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3941–3951, 2017.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

[5] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," *arXiv preprint arXiv:1808.00158*, 2018.

[6] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 681–687.

[7] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[8] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Advances in Neural Information Processing Systems*, 2014, pp. 2141–2149.

[9] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[10] O. K. Oyedotun, A. E. R. Shabayek, D. Aouada, and B. Ottersten, "Highway network block with gates constraints for training very deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1658–1667.

[11] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377–2385.

[12] "Bu-3d facial expression dataset," Aug. 2018. [Online]. Available: http://www.cs.binghamton.edu/ lijun/Research/3DFE/3DFE_Analysis.html

[13] "Washington object classification dataset," Jul. 2018. [Online]. Available: https://rgbd-dataset.cs.washington.edu

[14] O. K. Oyedotun, G. G. Demisse, A. E. R. Shabayek, D. Aouada, and B. E. Ottersten, "Facial expression recognition via joint deep learning of rgb-depth map latent representations." in *ICCV Workshops*, 2017, pp. 3161–3168.

[15] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. IEEE, 2006, pp. 211–216.

[16] X. Zhao, D. Huang, E. Dellandrea, and L. Chen, "Automatic 3d facial expression recognition based on a bayesian belief net and a statistical facial feature model," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3724–3727.

[17] H. Li, J. Sun, D. Wang, Z. Xu, and L. Chen, "Deep representation of facial geometric and photometric attributes for automatic 3d facial expression recognition," *arXiv preprint arXiv:1511.03015*, 2015.

[18] H. Tang and T. S. Huang, "3d facial expression recognition based on properties of line segments connecting facial feature points," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[19] H. Li, H. Ding, D. Huang, Y. Wang, X. Zhao, J.-M. Morvan, and L. Chen, "An efficient multimodal 2d+ 3d feature-based approach to automatic facial expression recognition," *Computer Vision and Image Understanding*, vol. 140, pp. 83–92, 2015.

[20] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3d facial expression recognition using geometric scattering representation," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–6.

[21] H. Li, J. Sun, D. Wang, Z. Xu, and L. Chen, "Deep representation of facial geometric and photometric attributes for automatic 3d facial expression recognition," *arXiv preprint arXiv:1511.03015*, 2015.